

Do Value-Added Estimates Add Value?

Accounting for Learning Dynamics

Tahir Andrabi
Jishnu Das
Asim Ijaz Khwaja
Tristan Zajonc

The World Bank
Development Research Group
Human Development and Public Services Team
September 2009



Abstract

Evaluations of educational programs commonly assume that what children learn persists over time. The authors compare learning in Pakistani public and private schools using dynamic panel methods that account for three key empirical challenges to widely used value-added models: imperfect persistence, unobserved student heterogeneity, and measurement error. Their estimates suggest that only a fifth to a half of learning persists between grades

and that private schools increase average achievement by 0.25 standard deviations each year. In contrast, estimates from commonly used value-added models significantly understate the impact of private schools' on student achievement and/or overstate persistence. These results have implications for program evaluation and value-added accountability system design.

This paper—a product of the Human Development and Public Services Team, Development Research Group—is part of a larger effort in the department to expand our knowledge of child learning and test scores as a broad measure of educational outcomes. Policy Research Working Papers are also posted on the Web at <http://econ.worldbank.org>. The author may be contacted at jdas1@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Do Value-Added Estimates Add Value? Accounting for Learning Dynamics*

Tahir Andrabi Jishnu Das Asim Ijaz Khwaja Tristan Zajonc[†]

1 Introduction

Models of learning often assume that children’s achievement persists between grades—what a child learns today largely stays with her tomorrow. Yet recent research highlights that treatment effects measured by test scores fade rapidly, both in randomized interventions and observational studies. Jacob, Lefgren and Sims (2008), Kane and Staiger (2008), and Rothstein (2008) find that teacher effects dissipate by between 50 and 80 percent over one year. The same pattern holds in several studies of supplemental education programs in developed and developing countries. Currie and Thomas (1995) document the rapid fade out of Head Start’s impact in the United States, and Glewwe, Ilias and Kremer (2003) and Banerjee et al. (2007) report on education experiments in Kenya and India where over 70 percent of the one-year treatment effect is lost after an additional year. Low persistence may in fact be the norm rather than the exception, and a central feature of learning.

Low persistence has critical implications for commonly used program evaluation strategies that rest heavily on assumptions about or estimation of persistence. Using primary data on public and private schools in Pakistan, this paper addresses the challenges to value-added evaluation strategies posed by 1) imperfect persistence of achievement, 2) heterogeneity in

*An earlier version of this paper also circulated under the title “Here Today, Gone Tomorrow? Examining the Extent and Implications of Low Persistence in Child Learning”.

[†]*tandrabi@pomona.edu*, Pomona College. *jdas1@worldbank.org*, World Bank, Washington DC and Center for Policy Research, New Delhi; *akhwaja@ksg.harvard.edu*, Kennedy School of Government, Harvard University, BREAD, NBER; *tzajonc@fas.harvard.edu*, Kennedy School of Government, Harvard University. We are grateful to Alberto Abadie, Chris Avery, David Deming, Pascaline Dupas, Brian Jacob, Dale Jorgenson, Elizabeth King, Karthik Muralidharan, David McKenzie, Rohini Pande, Lant Pritchett, Jesse Rothstein, Douglas Staiger, Tara Vishwanath, and seminar participants at Harvard, NEUDC and BREAD for helpful comments on drafts of this paper. This research was funded by grants from the Poverty and Social Impact Analysis and Knowledge for Change Program Trust Funds and the South Asia region of the World Bank. The findings, interpretations, and conclusions expressed here are those of the authors and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments they represent.

learning, and 3) measurement error in test scores. We find that ignoring any of these learning dynamics biases estimates of persistence and can dramatically affect estimates of the value-added of private schools.

To fix concepts, consider a simple model of learning evolution, $y_{it} = \alpha T_{it} + \beta y_{i,t-1} + \eta_i + v_{it}$, where y_{it} is child achievement measured by test scores in period t , T_{it} is the treatment or program effect in period t , and η_i is unobserved student ability that speeds learning each period. We refer to β , the parameter that links test scores across periods, as persistence.¹ The canonical restricted value-added or gain-score model assumes that $\beta = 1$ (for examples, see Hanushek (2003)). When $\beta < 1$, test scores exhibit mean reversion. Estimates of the treatment or program effect, α , that assume $\beta = 1$ will be biased if the baseline achievement of the treatment and control groups differs and persistence is imperfect. This has led many researchers to advocate leaving lagged achievement on the right-hand side. However doing so is not entirely straightforward: if estimated by OLS, omitted heterogeneity that speeds learning, η_i , will generally bias β upward and any measurement in test scores $y_{i,t-1}$ will bias β downward. Both the estimate of persistence β and the treatment effect α may remain biased when estimated by standard methods.

To address these concerns, we introduce techniques from the dynamic panel literature (Arellano and Honore, 2001; Arellano, 2003) that require three years of data. There are several findings. First, we find that learning persistence is low: only a fifth to a half of achievement persists between grades. That is, β is between 0.2 and 0.5 rather than closer to 1. These estimates are remarkably similar to those obtained in the United States (Jacob, Lefgren and Sims, 2008; Kane and Staiger, 2008; Rothstein, 2008). Second, OLS estimates of β are contaminated both by measurement error in test scores *and* unobserved student-level heterogeneity in learning. Ignoring both biases leads to higher persistence estimates between 0.5 and 0.6; correcting only for measurement error results in estimates between 0.7 and 0.8. For persistence, the upward bias from omitted heterogeneity outweighs measurement error attenuation.

Third, our estimate of the private schooling effect is highly sensitive to the persistence parameter. Since private schooling is a school input that is continually applied and leads to a large baseline gap in achievement, this is expected. Indeed, we find that incorrectly assuming $\beta = 1$ significantly understates and occasionally yields the wrong sign for private schools' impact on achievement—providing a compelling example of Lord's paradox (Lord, 1967). The restricted value-added model suggests that private schools contribute no more than public schools; in contrast, our dynamic panel estimates suggest large and significant

¹There are several different uses of the term persistence in the education literature. We refer to persistence as the fraction of knowledge that persist from one period to the next. The education literature, however, also uses the term “persistence” to indicate the probability of continuation from grade to grade (as opposed to dropping out), or to indicate a child's motivation or propensity to complete tasks in the face of adversity.

contributions ranging from 0.19 to 0.32 standard-deviations a year.² Notably, the lagged value-added model estimated by OLS gives similar results for the private school effect as our more data intensive dynamic panel methods. This is due to the countervailing heterogeneity and measurement error biases on β and because lagged achievement can act as a partial proxy for omitted heterogeneity in learning.³

Finally, towards an economic interpretation of low persistence, we use question-level exam responses as well as household expenditure and time-use data to explore whether psychometric testing issues, behavioral responses, or forgetting contribute to low persistence—causes that have different welfare implications. This investigation suggests that measurement error, mechanical psychometric testing, and behavioral response based explanations are insufficient. Understanding the behavioral or technological reasons for low persistence remains a critical issue in the literature.

The value-added of our contribution is several fold. To begin with, we show that the restricted value-added estimates based on longitudinal data may be *worse* than the naive cross-sectional OLS estimates. Second, we demonstrate how unobserved heterogeneity in learning and measurement error in test scores can bias estimates of persistence. The low persistence we find implies that long-run extrapolations from short-run impacts are fraught with danger. In the model above, the long-run impact of continued treatment is $\alpha/(1 - \beta)$; with estimates of β around 0.2 to 0.5, these gains may be much smaller than those obtained by assuming that β is close to 1.⁴ Third, we find that students in Pakistan’s private schools learn significantly more each year than their public school counterparts but that the popular gain-score or restricted value-added model would have detected no effect of private schools on learning.⁵ From a public

²Harris and Sass (2006) find that the persistence parameter makes little difference to estimates of teacher effects, while we find it starkly affects the estimates of school type. This can be explained by the relative gaps in baseline achievement. It is likely that a child does not continue with the same teacher, or an equally good teacher, over time. Hence, even if we don’t observe children’s educational history, two children who currently have different teachers may have been exposed to a similar quality teachers in the past. As such, children with different teachers often do not differ substantially in their baseline learning levels. In contrast, given that there is little switching across school types, children currently in different schools differ substantially in baseline learning levels.

³This results suggests that correcting for measurement error alone may do more harm than good. For example, Ladd and Walsh (2002) correct for measurement error in the lagged value-added model of school effects by instrumenting using double-lagged test scores but don’t address potential omitted heterogeneity. They show this correction significantly changes school rankings and benefits poorly performing districts. Given that we find unobserved heterogeneity in learning rates, rankings that correct for measurement error may be poorer than those that do not.

⁴For example, Krueger and Whitmore (2001), Angrist et al. (2002), Krueger (2003), and Gordon, Kane and Staiger (2006) calculate the economic return of various educational interventions by citing research linking test scores to earnings of young adults (e.g. Murnane, Willett and Levy, 1995; Neal and Johnson, 1996). Although effects on learning as measured by test-scores may fade, non-cognitive skills that are rewarded in the labor market could persist. For instance, Currie and Thomas (1995), Deming (2008) and Schweinhart et al. (2005) provide evidence of long run effects of Head Start and the Perry Preschool Project, even though cognitive gains largely fade after children enroll in regular classes.

⁵An alternative identification strategy to value-added models that we do not pursue here is instrumental

finance point of view, these different estimates matter particularly since per pupil expenditures are lower in private relative to public schools.⁶ Our results are consistent with growing evidence that relatively inexpensive, mainstream, private schools hold potential in the developing country context (Jimenez, Lockheed and Paqueo, 1991; Alderman, Orazem and Paterno, 2001; Angrist et al., 2002; Alderman, Kim and Orazem, 2003; Tooley and Dixon, 2003; Muralidharan and Kremer, forthcoming; Andrabi, Das and Khwaja, 2008). Overall, our general results support a movement towards long-run, experimental evaluations of educational interventions.

A final contribution of our work is that it applies a wider set of econometric tools from the dynamic panel data literature than have been typically used in the education literature. In the use of dynamic panel methods, our estimators bear greatest resemblance to those discussed by Schwartz and Zabel (2005) and Sass (2006). Both use simple dynamic panel estimators, in the first case using school-level data and in the second using the Arellano and Bond (1991) differences GMM approach. Santibanez (2006) also uses the Arellano and Bond (1991) estimator to analyze the impact of teacher quality. Our efforts extend to include system GMM approaches (Arellano and Bover, 1995) and to address measurement error in test scores and alternative assumptions regarding omitted heterogeneity. We find both are important.

The rest of the paper is organized as follows: Section 2 presents the basic education production function analogy and discusses the specification and estimation of the value-added approximations to it. Section 3 summarizes our data. Section 4 reports our main results, several robustness checks, and a preliminary exploration of the economic interpretation of persistence. Section 5 concludes by discussing implications for experimental and non-experimental program evaluation.

2 Empirical Learning Framework

The “education production function” approach to learning relates current achievement to all previous inputs. Boardman and Murnane (1979) and Todd and Wolpin (2003) provide two accounts of this approach and the assumptions it requires; the following is a brief summary.⁷ Using notation consistent with the dynamic panel literature, we aggregate all inputs into a single vector \mathbf{x}_{it} and exclude interactions between past and present inputs. Achievement for

variables. We are exploring such strategies using plausible exogenous geographical variation in a companion paper focused on the returns to private school education in Pakistan and competition between public and private schools. Our emphasis here is on the challenges faced by popular value-added strategies.

⁶For details on the costs of private schooling in Pakistan see Andrabi, Das and Khwaja (2008).

⁷Researchers generally assume that the model is additively separable across time and that input interactions can be captured by separable linear interactions. Cunha, Heckman and Schennach (2006) and Cunha and Heckman (2007) are two exceptions to this pattern, where dynamic complementarity between early and late investments and between cognitive and non-cognitive skills are permitted.

child i at time (grade) t is therefore

$$y_{it}^* = \alpha_1' \mathbf{x}_{it} + \alpha_2' \mathbf{x}_{i,t-1} + \dots + \alpha_t' \mathbf{x}_{i1} + \sum_{s=1}^{s=t} \theta_{t+1-s} \mu_{is}, \quad (1)$$

where y_{it}^* is true achievement, measured without error, and the summed μ_{is} are cumulative productivity shocks.⁸ Estimating (1) is generally impossible because researchers do not observe the full set of inputs, past and present. The value-added strategy makes estimation feasible by rewriting (1) to avoid the need for past inputs. Adding and subtracting βy_{it}^* , normalizing θ_1 to unity, and assuming that coefficients decline geometrically ($\alpha_j = \beta \alpha_{j-1}$ and $\theta_j = \beta \theta_{j-1}$ for all j) yields the *lagged value-added model*

$$y_{it}^* = \alpha_1' \mathbf{x}_{it} + \beta y_{i,t-1}^* + \mu_{it}. \quad (2)$$

The basic idea behind this specification is that lagged achievement will capture the contribution of all previous inputs and any past unobservable endowments or shocks. As before, we refer to α as the *input coefficient* and β as the *persistence coefficient*. Finally, imposing the restriction that $\beta = 1$ yields the *gain-score* or *restricted value-added model* that is often used in the education literature:

$$y_{it}^* - y_{i,t-1}^* = \alpha_1' \mathbf{x}_{it} + \mu_{it}.$$

This model asserts that past achievement contains no information about future gains, or equivalently, that an input's effect on any subsequent level of achievement does not depend on how long ago it was applied. As we will see from our results, the assumption that $\beta = 1$ is clearly violated in the data, and increasingly it appears, in the literature, as well. As a result, we will focus primarily on estimating (2).

There are two potential problems with estimating (2). First, the error term μ_{it} could include individual (child-level) heterogeneity in *learning* (e.g., $\mu_{it} \equiv \eta_i + v_{it}$). Lagged achievement only captures individual heterogeneity if it enters through a one-time process or endowment, but talented children may also *learn* faster. Since this unobserved heterogeneity enters in each period, $\text{Cov}(y_{i,t-1}^*, \mu_{it}) > 0$ and β will be biased upwards.

The second likely problem is that test scores are inherently a noisy measure of latent achievement. Letting $y_{it} = y_{it}^* + \varepsilon_{it}$ denote observed achievement, we can rewrite the latent lagged value-added model (2) in terms of observables. The full error term now includes measurement

⁸This starting point is more restrictive than the more general starting framework presented by Todd and Wolpin (2003). In particular, it assumes an input applied in first grade has the same effect on first grade scores as an input applied in second grade has on second grade scores.

error, $\mu_{it} + \varepsilon_{it} - \beta\varepsilon_{i,t-1}$.

Dropping all the inputs to focus solely on the persistence coefficient, the expected bias due to both of these sources is

$$\text{plim } \beta_{OLS} = \beta + \left(\frac{\text{Cov}(\eta_i, y_{i,t-1}^*)}{\sigma_{y^*}^2 + \sigma_\varepsilon^2} \right) - \left(\frac{\sigma_\varepsilon^2}{\sigma_{y^*}^2 + \sigma_\varepsilon^2} \right) \beta. \quad (3)$$

The coefficient is biased upward by learning heterogeneity and downward by measurement error. These effects only cancel exactly when $\text{Cov}(\eta_i, y_{i,t-1}^*) = \sigma_\varepsilon^2 \beta$ (Arellano, 2003).

Furthermore, bias in the persistence coefficient leads to bias in the input coefficients, α . To see this, consider imposing a biased $\hat{\beta}$ and estimating the resulting model

$$y_{it} - \hat{\beta}y_{i,t-1} = \alpha' \mathbf{x}_{it} + [(\beta - \hat{\beta})y_{i,t-1} + \mu_{it} + \varepsilon_{it} - \beta\varepsilon_{i,t-1}].$$

The error term now includes $(\beta - \hat{\beta})y_{i,t-1}$. Since inputs and lagged achievement are generally positively correlated, the input coefficient will, in general, be biased downward if $\hat{\beta} > \beta$. The precise bias, however, depends on the degree of serial correlation of inputs and on the potential correlation between inputs and learning heterogeneity that remains in μ_{it} .

This is more clearly illustrated in the case of the restricted value-added model (assuming that $\beta = 1$) where:

$$\text{plim } \hat{\alpha}_{OLS} = \alpha - (1 - \beta) \frac{\text{Cov}(\mathbf{x}_{it}, y_{i,t-1})}{\text{Var}(\mathbf{x}_{it})} + \frac{\text{Cov}(\mathbf{x}_{it}, \eta_i)}{\text{Var}(\mathbf{x}_{it})}. \quad (4)$$

Therefore, if indeed there is perfect persistence as assumed and if inputs are uncorrelated with η_i , OLS yields consistent estimates of the parameters α . However, if $\beta < 1$, OLS estimation of α now results in two competing biases. By assuming an incorrect persistence coefficient we leave a portion of past achievement in the error term. This misspecification biases the input coefficient downward by the first term in (4). The second term captures possible correlation between current inputs and omitted learning heterogeneity. If there is none, then the second term is zero, and the bias will be unambiguously negative.

2.1 Addressing Child-Level Heterogeneity: Dynamic Panel Approaches to the Education Production Function

Dynamic panel approaches can address omitted child-level heterogeneity in value-added approximations of the education production function. We interpret the value-added model (2) as an autoregressive dynamic panel model with unobserved student-level effects:

$$y_{it}^* = \alpha' \mathbf{x}_{it} + \beta y_{i,t-1}^* + \mu_{it}, \quad (5)$$

$$\mu_{it} \equiv \eta_i + v_{it}. \quad (6)$$

Identification of β and α is achieved by imposing appropriate moment conditions. Following Arellano and Bond (1991) and Arellano and Bover (1995), we focus on linear moment conditions and split our analysis into three groups: “differences” GMM, “differences and levels” GMM, and “levels only” GMM, which respectively refer to whether the estimates are based on the undifferenced “levels” equation (5), a differenced equation (see equation (7) below), or both. The section below provides a brief overview of the estimators we explore. Table 1 summarizes the estimators, including the standard static value-added estimators (M1-M4) and dynamic panel estimators (M5-M10). For more complete descriptions, Arellano and Honore (2001) and Arellano (2003) provide excellent reviews of these and other panel models.

2.1.1 Differences GMM: Switching estimators

As noted previously, the value-added model differences out omitted endowments that might be correlated with the inputs. It does not, however, difference out heterogeneity that speeds learning. To accomplish this, the basic intuition behind the Arellano and Bond (1991) difference GMM estimator is to difference again. Differencing the dynamic panel specification of the lagged value-added model (5) yields

$$y_{it}^* - y_{i,t-1}^* = \alpha' (\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) + \beta (y_{i,t-1}^* - y_{i,t-2}^*) + [v_{it} - v_{i,t-1}]. \quad (7)$$

Here, the differenced model eliminates the unobserved fixed effect η_i . However, (7) cannot be estimated by OLS because $y_{i,t-1}^*$ is correlated by construction with $v_{i,t-1}$ in the error term. Arellano and Bond (1991) propose instrumenting for $y_{i,t-1}^* - y_{i,t-2}^*$ using lags two periods and beyond, such as $y_{i,t-2}^*$, or certain inputs, depending on the exogeneity conditions. These lags are uncorrelated with the error term but are correlated with the change in lagged achievement, provided $\beta < 1$. The input coefficient, in our case the added contribution of private schools, is primarily identified from the set of children who switch schools in the observation period.

The implementation of the difference GMM approach depends on the precise assumptions

about inputs. We consider two candidate assumptions: strictly exogenous inputs (M5) and predetermined inputs (M6). Strict exogeneity assumes past disturbances do not affect current and future inputs, ruling out feedback effects. In the educational context, this is a strong assumption. A child who experiences a positive or negative shock may adjust inputs in response. In our case, a shock may cause a child to switch schools.

To account for this possibility, we also consider the weaker case where inputs are predetermined but not strictly exogenous. Specifically, the predetermined inputs case assumes that inputs are uncorrelated with present and future disturbances but are potentially correlated with past disturbances. This case also assumes lagged achievement is uncorrelated with present and future disturbances. Compared to strict exogeneity, this approach uses only lagged inputs as instruments. Switching schools is instrumented by the original school type, allowing switches to depend on previous shocks. This estimator remains consistent if a child switches school at the same time as an achievement shock but still rules out parents anticipating and adjusting to future expected shocks.

2.1.2 Levels and differences GMM: Uncorrelated or constantly correlated effects

One difficulty with the differences GMM approach (M5 and M6) is that time-invariant inputs drop out of the estimated equation and their effects are therefore not identified. In our case, this means that the identification of the private school effect is based on the five percent of children who switch between public and private schools. We address the limited time-series variation using the levels and differences GMM framework proposed by Arellano and Bover (1995) and extended by Blundell and Bond (1998). Levels and differences GMM estimates a system of equations, one for the undifferenced levels equation (5) and another for the differenced equation (7). Further assumptions regarding the correlation between inputs and heterogeneity (though not necessarily between heterogeneity and lagged achievement) yield additional instruments.

We first consider predetermined inputs that have a constant correlation with the individual effects (M7). While inputs may be correlated with the omitted effects, constant correlation implies switching is not. The constant correlation assumption implies that $\Delta \mathbf{x}_{it}$ are available as instruments in the levels equation (Arellano and Bover, 1995). In the context of estimating school type, this estimator can be viewed as a levels and differences switching estimator since it relies on children switching school types in both the levels and differences equations. In practice, we often must assume that any time-invariant inputs are uncorrelated with the fixed effect or the levels equation, which includes the time-invariant inputs, is not fully identified.

A second possibility is that inputs are predetermined but are also uncorrelated with the omitted effects (M8). This allows using inputs \mathbf{x}_i^t as instruments in the levels model (5). The required assumption is fairly strong; it is natural to believe that inputs are correlated with

the omitted effect. Certainly, the decision to attend private school may be correlated with the child's ability to learn. At the same time, the assumption is weaker than OLS estimation of lagged value-added model since the model (M8) allows for the omitted fixed effect to be correlated with lagged achievement.

2.1.3 Levels GMM: Conditional mean stationarity

In some instances, it may be reasonable to assume that, while learning heterogeneity exists, it does not affect achievement gains. A talented child may be so far ahead that imperfect persistence cancels the benefit of faster learning. That is, individual heterogeneity may be uncorrelated with gains, $y_{it}^* - y_{it-1}^*$, but not necessarily with *learning*, $y_{it}^* - \beta y_{it-1}^*$. This situation arises when the initial conditions have reached a convergent level with respect to the fixed effect such that

$$y_{i1}^* = \frac{\eta_i}{1 - \beta} + d_i, \quad (8)$$

where $t = 1$ is the first observed period and not the first period in the learning life-cycle. Blundell and Bond (1998) discuss this type of conditional mean stationarity restriction in considerable depth. As they point out, the key assumption is that initial deviations, d_i , are uncorrelated with the level of $\eta_i/(1 - \beta)$. It does not imply that the achievement path, $\{y_{i1}^*, y_{i2}^*, \dots, y_{iT}^*\}$, is stationary; inputs, including time dummies, continue to spur achievement and can be nonstationary. The assumption only requires that, conditional on the full set of controls and common time dummies, the individual effect does not influence achievement gains.

While this assumption seems too strong in the context of education, we discuss it because the dynamic panel literature has documented large downward biases of other estimators when the instruments are weak (e.g. Blundell and Bond, 1998). This occurs when persistence is perfect ($\beta = 1$) since the lagged value-added model then exhibits a unit root and lagged tests scores become weak instruments in the differenced model. The conditional mean stationarity assumption provides an additional $T - 2$ non-redundant moment conditions that can augment the system GMM estimators. While a fully efficient approach uses these additional moments along with typical moments in the differenced equation, the conditional mean stationarity assumption ensures strong instruments in the levels equation to identify β . Thus, if we prefer simplicity over efficiency, we can estimate the model using levels GMM or 2SLS and avoid the need to use a system estimator. In this simpler approach, we instrument the undifferenced value-added model (5) using lagged changes in achievement, Δy_i^{*t-1} , and either changes in inputs, $\Delta \mathbf{x}_i^t$, or inputs directly, \mathbf{x}_i^t , depending on whether inputs are constantly correlated (M9) or are uncorrelated with the individual effect (M10).

2.2 Addressing Measurement Error in Test Scores

Measurement error in test scores is a central feature of educational program evaluation. Ladd and Walsh (2002), Kane and Staiger (2002), and Chay, McEwan and Urquiola (2005) all document how test-score measurement error can pose difficulties for program evaluation and value-added accountability systems. In the context of value-added estimation, measurement error attenuates the coefficient on lagged achievement and can bias the input coefficient in the process. Dynamic panel estimators do not address measurement error on their own. For instance, if we replace true achievement with observed achievement in the standard Arellano and Bond (1991) setup, (7) becomes

$$\Delta y_{it} = \boldsymbol{\alpha}' \Delta \mathbf{x}_{it} + \beta \Delta y_{i,t-1} + [\Delta v_{it} + \Delta \varepsilon_{i,t} - \beta \Delta \varepsilon_{i,t-1}]. \quad (9)$$

The standard potential instrument, $y_{i,t-2}$, is uncorrelated with Δv_{it} but is correlated with $\Delta \varepsilon_{i,t-1} = \varepsilon_{i,t-1} - \varepsilon_{i,t-2}$ by construction.

The easiest solution is to use either three-period lagged test scores or alternate subjects as instruments. In the dynamic panel models discussed above, correcting for measurement error using additional lags requires four years of data for each child—a difficult requirement in most longitudinal datasets, including ours. We therefore use alternate subjects, although doing so does not address the possibility of correlated measurement error across subjects.

An alternative to instrumental variables strategies is to correct for measurement error analytically using the standard error of each test score, available from Item Response Theory.⁹ Because the standard error is heteroscedastic—tests discriminate poorly between children at the tails of the ability distribution—one can gain efficiency by using the heteroscedastic errors-in-variables (HEIV) procedure outlined in Sullivan (2001) and followed by Jacob and Lefgren (2005), among others. Appendix A provides a detailed explanation of this analytical correction. While this correction is easy to apply in an OLS model, it becomes considerably more complicated in the dynamic panel context, and we therefore use an instrumental variable strategy for most of our estimators.

3 Data

To demonstrate these issues, we use data collected by the authors as part of the Learning and Educational Achievement in Punjab Schools (LEAPS) project, an ongoing survey of learning in Pakistan. The sample comprises 112 villages in 3 districts of Punjab: Attock, Faisalabad, and Rahim Yar Khan. Because the project was envisioned in part to study the dramatic rise of

⁹Item Response Theory provides the standard error for each score from the inverse Fisher information matrix after ML estimation of the IRT model. This standard error is reported in many educational datasets.

private schools in Pakistan, the 112 villages in these districts were chosen randomly from the list of all villages with an existing private school. As would be expected given the presence of a private school, the sample villages are generally larger, wealthier, and more educated than the average rural village. Nevertheless, at the time of the survey, more than 50 percent of the province’s population resided in such villages (Andrabi, Das and Khwaja, 2006).

The survey covers all schools within the sample village boundaries *and* within a short walk of any village household. Including schools that opened and closed over the three rounds, 858 schools were surveyed, while three refused to cooperate. Sample schools account for over 90 percent of enrollment in the sample villages.

The first panel of children consists of 13,735 third-graders, 12,110 of which were tested in Urdu, English, and mathematics. These children were subsequently followed for two years and retested in each period. Every effort was made to track children across rounds, even when they were not promoted. In total, 12 percent and 13 percent of children dropped out or were lost between rounds one and two, and two and three, respectively.¹⁰ In addition to being tested, 6,379 children—up to ten in each school—were randomly administered a survey including anthropometrics (height and weight) and detailed family characteristics such parental education and wealth, as measured by principal components analysis analysis of 20 assets. When exploring the economic interpretation of persistence, we also use a small subsample of approximately 650 children that can be matched to a detailed household survey that includes, among other things, child and parental time use and educational spending.

For our analysis, we use two subsamples of the data: all children who were tested in all three years (N=8120) and children who were tested *and* given a detailed child survey in all three years (N=4031). Table 2 presents the characteristics of these children split by whether they attend public or private schools. The patterns across each subsample is relatively stable. Children attending private schools are slightly younger, have fewer elder siblings, and come from wealthier and more educated households. Years of schooling, which largely captures grade retention, is lower in private schools. Children in private schools are also less likely to have a father living at home, perhaps due to a migration or remittance effect on private school attendance.

The measures of achievement are based on exams in English, Urdu (the vernacular), and mathematics. The tests were relatively long (over 40 questions per subject) and were designed to maximize the precision over a range of abilities in each grade. While a fraction of questions

¹⁰Attrition in private schools is 2 percentage points higher than in public schools. Children who drop out between rounds one and two have scores roughly 0.2 s.d. lower than children that don’t. Controlling for school type and drop out status, drop outs in private schools are slightly better (0.05 sd) than children in public schools, although the difference is only statistically significant for math. Given the small relative differences in attrition between public and private schools, additional corrections for attrition are unlikely to significantly affect our results.

changed over the years, the content covered remained consistent, and a significant portion of questions appeared across all years. To avoid the possibility of cheating, the tests were administered directly by our project staff and not by classroom teachers. The tests were scored and equated across years by the authors using Item Response Theory so that the scale has cardinal meaning. Preserving cardinality is important for longitudinal analysis since many other transformations, such as the percent correct score or percentile rank, are bounded artificially by the transformations that describe them. By comparison, IRT scores attempt to ensure that change in one part of the distribution is equal to a change in another, in terms of the latent trait captured by the test. Children were tested in third, fourth, and fifth grades during the winter at roughly one year intervals. Because the school year ends in the early spring, the test scores gains from third to fourth grade are largely attributable to the fourth grade school.

4 Results

4.1 Cross-sectional and Graphical Results

Before presenting our estimates of learning persistence and the implied private school effect, we provide some rough evidence for a significant private school effect using cross-sectional and graphical evidence. These results do not take advantage of the more sophisticated specifications above but nevertheless provide initial evidence that the value-added of private schools is large and significant.

4.1.1 Baseline estimates from cross-section data

Table 3 presents results for a cross-section regression of third grade achievement on child, household, and school characteristics. These regressions provide some initial evidence that the public-private gap is more than omitted variables and selection. Adding a comprehensive set of child and family controls reduces the estimated coefficient on private schools only slightly. Adding village fixed effects also does not change the coefficient, even though the R^2 increases substantially. Across all baseline specifications, the gap remains large: over 0.9 standard deviations in English, 0.5 standard deviations in Urdu, and 0.4 standard deviations in mathematics.

Besides the coefficient on school type, few controls are strongly associated with achievement. By far the largest other effect is for females, who outperform their male peers in English and Urdu. However, even for Urdu, where the female effect is largest, the private school effect is still nearly three times as large. Height, assets, and whether the father (and for Column 3, mother) is educated past elementary school also enter the regression as positive and significant. More elder brothers and more years of schooling (i.e. being previously retained) correlates with lower achievement. Children with a mother living at home perform worse although this result is driven

by an abnormal subpopulation of two percent of children with absent mothers. Overall, these results confirm mild positive selection into private schools but also suggest that, controlling a host of other observables typically not available in other datasets (such as child height and household assets) does not alter significantly the size of the private schooling coefficient.

4.1.2 Graphical evidence

Figure 1 plots learning levels in the tested subjects (English, mathematics, and the vernacular, Urdu) over three years. While, levels are always higher for children in private schools, there is little difference in learning gains (the gradient) between public and private schools. This illustrates why a specification that uses learning gains (i.e., assumes perfect persistence) would conclude that private schools add no greater value to learning than their public counterparts.

Many of the dynamic panel estimators that we explore identify the private school effect using children who switch schools. Figure 2 illustrates the patterns of achievement for these children. For each subject we plot two panels: the first containing children who start in public school and the second containing those who start in private school. We then graph achievement patterns for children who never switch, switch after third grade, and switch after fourth grade. For simplicity, we exclude children who switch back and forth between school types.

As the table at the bottom of the figure shows, very few children change schools. Only 48 children move from public to private schools in fourth grade, while 40 move in fifth grade. Consistent with the role of private schools serving primarily younger children, 167 children switch to public schools in fourth grade, and 160 switch in fifth grade. These numbers are roughly double the number of children available for our estimates that include controls, since only a random subset of children were surveyed regarding their family characteristics.

Even given the small number of children switching school types, Figure 2 provides preliminary evidence that the private school effect is not simply a cross-sectional phenomenon. In all three subjects, children who switch to private schools between third and fourth grade experience large achievement gains. Children switching from private schools to public schools exhibit similar achievement patterns, except reversed. Moving to a public school is associated with slower learning or even learning losses. Most gains or losses occur immediately after moving; once achievement converges to the new level, children experience parallel growth in public and private schools.

4.2 OLS and Dynamic Panel Value-Added Estimates

Tables 4 (English), 5 (Urdu), and 6 (mathematics) summarize our main value-added results. All estimates include the full set of controls in the child survey sample, the survey date, round (grade) dummies, and village fixed effects. For brevity, we only report the persistence and

private school coefficients.¹¹ We group the discussion of our results in three domains: estimates of the persistence coefficient, estimates of the private schooling coefficient, and regression diagnostics.

4.2.1 The persistence parameter

We immediately reject the hypothesis of perfect persistence ($\beta = 1$). Across all specifications and all subject (except M1 which imposes $\beta = 1$), the estimated persistence coefficient is significantly lower than one, even in the specifications that correct for measurement error only and should be biased upward (M3 and M4). The typical lagged value-added model (M2), which assumes no omitted student heterogeneity and no measurement error, returns estimates between 0.52 and 0.58 for the persistence coefficient. Correcting only for measurement error by instrumenting using the two alternate subjects (M3), or using the analytical correction described by in the appendix (M4), increases the persistence coefficient to between 0.70 and 0.79, *consistent with significant measurement error attenuation*. This estimate, however, remains biased upward by omitted heterogeneity.

Moving to our dynamic panel estimators, Panel B of each table gives the Arellano and Bond (1991) difference GMM estimates under the assumption that inputs are strictly exogenous (M5) or predetermined (M6). In English and Urdu, the persistence parameter falls to between 0.19 and 0.35. The estimates are (statistically) different from models that correct for measurement error only. *In other words, omitted heterogeneity in learning exists, and biases the static estimates upward*. For mathematics, the estimated persistence coefficient is indistinguishable from zero, considerably below all the other estimates. These estimates are higher and somewhat more stable in the systems GMM approach summarized in Panel C (M7, M8).

With the addition of a conditional mean stationarity assumption (Panel D), we can more precisely estimate the persistence coefficient. In this model, we only use moments in levels to illustrate a dynamic panel estimator that improves over the lagged value-added model estimated by OLS but doesn't require estimating a system of equations. The persistence coefficient rises substantially to between 0.39 and 0.56. This upward movement is consistent with a violation of the stationarity assumption (the fixed-effect still contributes to achievement growth) but an overall reduction in the omitted heterogeneity bias. Across the various dynamic panel models and subjects, estimates of the persistence parameter vary from 0.2 to 0.55. However the highest dynamic panel estimates come from assuming conditional mean stationary, which is likely too strong in the context of education.

¹¹As discussed, time-invariant controls drop out of the differenced models. For the system and levels estimators we also assume, by necessity, that time-invariant controls are uncorrelated with the fixed effect or act as proxy variables.

4.2.2 The contribution of private schools

Assuming perfect persistence biases the private school coefficient downward. For English, the estimated private school effect in the restricted model that incorrectly assumes $\beta = 1$ (Panel A, Table 4) is negative and significant. For Urdu and mathematics, the private school coefficient is small and insignificant or marginally significant (Panel A, Tables 5 and 6). By comparison, all the dynamic panel estimates are positive and statistically significant, with the exception of one of the difference GMM estimates, which is too weak to identify the private school effect with any precision.

Panel C (levels and differences GMM) illustrates the benefit of a systems approach. Adding a levels equation (Panel C, Tables 4-6), using the assumption that inputs are constantly correlated or uncorrelated with the omitted effects, reduces the standard errors for the private school coefficient while maintaining the assumption that inputs are predetermined but not strictly exogenous. Under the scenario that private school enrollment is constantly correlated with the omitted effect (M7), the private school coefficient is large: 0.19 to 0.32 standard deviations depending on the subject and statistically significant. This estimate allows for past achievement shocks to affect enrollment decisions but assumes that switching school type is uncorrelated with unobserved student heterogeneity. This is our preferred estimate.

An overarching theme in this analysis is that the persistence parameter influences the estimated private school effect but that it is rarely possible to get enough precision to distinguish estimates based on different exogeneity conditions. This is largely due to the small number of children switching between public and private schools in our sample. In Figure 3, we graph the relationship between both coefficients explicitly. Rather than estimating the persistence coefficient, we assume a specific rate and then estimate the value-added model. That is, we use $y_{it} - \beta y_{i,t-1}$ as the dependent variable. This provides a robustness check for any estimated effects, requires only two years of data, and eliminates the need for complicated measurement error corrections. (It assumes, however, that inputs are uncorrelated with the omitted learning heterogeneity.) As expected given the large baseline differences, the estimated private school effect strongly depends on the assumed persistence rate. Moving from the restricted value-added model ($\beta = 1$) to the pooled cross-section model ($\beta = 0$) increases the estimated effect from negative or insignificant to large and significant. For most of the range of the persistence parameter, the private school effect is positive and significant, but pinning down the precise yearly contribution of private schooling depends on our assumptions about how children learn.

A couple of natural questions are how these estimates compare to the private-public differences reported in the cross-section and why the trajectories in Figure 1 are parallel even though the private school effect is positive. Controlling for observables suggests that after three years, children in private schools are 0.9 (English), 0.5 (Urdu), and 0.45 (mathematics) stan-

dard deviations ahead of their public school counterparts. If persistence is 0.4 and the yearly private school effect is 0.3, children’s trajectories will become parallel when that achievement gap reaches 0.5 ($= 0.3/(1 - 0.4)$). This is roughly the gap we find in Urdu and mathematics. Any small disagreement, including the larger gap in English, may be attributable to baseline selection effects. Thus our results can consistently explain the large baseline gap in achievement, the parallel achievement trajectories in public and private schools, and the significant and ongoing positive private school effect.

4.2.3 Regression diagnostics

For many of the GMM estimates, Hansen’s J test rejects the overidentifying restrictions implied by the model. This is troubling but not entirely unexpected. Different instruments may be identifying different local average treatment effects in the education context. For example, the portion of third grade achievement that remains correlated with fourth grade achievement may decay at a different rate than what was learned most recently. This is particularly true in an optimizing model of skill formation where parents smooth away shocks to achievement. In such a model, unexpected shocks to achievement, beyond measurement error, would fade more quickly than expected gains. Instrumenting using contemporaneous alternate subject scores will therefore more likely identify different parameters than instrumenting using previous year scores. Likewise, instrumenting using alternate lags and differenced achievement and inputs may also identify different effects. This type of heterogeneity is important and suggests that a richer model than a constant coefficient lagged value-added may be warranted.¹² Given the rejection of the overidentifying restrictions in some cases, the next section provides a series of robustness exercises around the estimation of the persistence parameter.

4.3 Robustness Checks

If our estimates are interpreted as forgetting, children lose over half of their achievement in a single year. For some subjects, such as mathematics, this fraction may be even larger. While the estimates reported may appear to be implausibly high, they match recent work on fade-out in value-added models, as well as the rapid fade-out observed in most educational interventions.

Table 7 summarizes six randomized (or quasi-randomized) interventions that followed children after the program ended. This follow-up enables estimation of both immediate and extended treatment effects. For the interventions summarized, the extended treatment effect represents test scores roughly one year after the particular program ended. For a number of

¹²Another common strategy to address potentially invalid instruments is to slowly reduce the instrument set, testing each subset, until the overidentification test is accepted or the model becomes just identified. We explored this approach but no clear story emerged. One result of note is that dropping the overidentifying inputs typically raises the the persistence coefficient slightly, to roughly 0.25 for math.

the interventions, the persistence coefficient is less than 0.10. In two interventions—learning incentives and grade retention—the coefficient is between 0.6 and 0.7. However, this higher level of persistence may in part be explained by the specific nature of these interventions.¹³ Although the link between fade-out in experimental studies and the persistence parameter is not always exact, all the evidence suggests that current learning does not carry over to future learning without loss and, in fact, these losses may be substantial.

Exploring the magnitude of the *potential* bias in a basic lagged value-added model can also give us some sense for whether our estimates are reasonable. Consider, for example, the bias in the regression $y_{it} = \alpha + \beta y_{i,t-1} + \eta_{it}$, where we have omitted all potential inputs and corrected only for measurement error bias. Our estimates of this model suggest that the persistence coefficient is at most 0.8 to 0.9—far higher than our highest dynamic panel estimates of around 0.5. Is this discrepancy reasonable?

Aggregating all the omitted contemporaneous inputs into one variable η_{it} implies the upward bias of the persistence coefficient is $\text{Cov}(\eta_{it}, y_{i,t-1}) / \text{Var}(y_{i,t-1})$. If the correlation between inputs in any two periods is a constant ρ_X , and all children in grade zero start from the same place, the persistence coefficient in a lagged value-added model for fourth grade will be biased upward to

$$\frac{\text{Cov}(\eta_{i4}, y_{i3})}{\text{Var}(y_{i3})} = \frac{\rho_X}{2\beta\rho_X - \beta + \beta^2 + 1}. \quad (10)$$

Figure 4 gives a graphical representation of this bias calculation. To read the graph, choose a true persistence coefficient, β (the dotted lines), and a degree of correlation of inputs over time, ρ_X (the horizontal axis). Given these choices, the y-axis reveals the persistence coefficient that a lagged value-added specification estimated by OLS would yield. Working with our estimates, if the true persistence effect, β , is 0.4 and inputs are correlated only 0.6 over time, the (incorrectly) estimated β will be 0.9. Given that the vast majority of inputs are fixed, this seems quite reasonable, and perhaps even too low.¹⁴

¹³In the case of grade retention, there is no real “post treatment” period since children always remain one grade behind after being retained. If one views grade retention as an ongoing multi-period treatment, then lasting effects can be consistent with low persistence. In the case of learning incentives, Kremer, Miguel and Thornton (2003) argue that student incentives increased effort (not just achievement) even after the program ended, leading to ongoing learning.

¹⁴Another way to get at the reasonableness of rapid fade out is motivated by Altonji, Elder and Taber’s (2005) assumption of equal selection on observed and unobserved variables. Absent controls and correcting for measurement error the persistence coefficient is 0.91, while the R^2 of the regression is 0.52. Adding controls raises the R^2 only modestly to 0.56 but at the same time reduces the estimated persistence coefficient to 0.74. Thus, just by explaining an additional four percent of the total variation, we reduced the persistence coefficient substantially. Assuming equal selection on observed and unobserved variables would lead to a persistence estimate *below* our dynamic panel estimates.

4.4 Why Is Persistence So Low?

The low estimates of persistence are worrying, not only for program evaluation but for the more substantive issue of how to improve cognitive achievement. One major concern is that imperfect persistence is a psychometric testing issue, and therefore not a “true” feature of the learning dynamic. For instance, later test forms may capture fundamentally different latent traits than earlier test forms. To address this concern, we replicated our results using IRT scores based solely on a common set of items that appeared on every test form—our tests had a significant number of overlapping items in each year. Our results were similar using these scores, with the difference GMM persistence estimates in fact dropping slightly. The score equating methods used to create a single cardinal measure of learning therefore do not appear to be the driving force behind the low observed persistence.

Several other possible mechanical explanations for low persistence are also unlikely. First, artificial ceiling effects can appear like low persistence in models that use bounded scores. To address this concern, we exclusively use unbounded IRT scale scores and our exam is designed to maximize the variation over the entire range of observed abilities. Second, cheating, often driven by high stakes testing, can create artificially low estimates of persistence. Jacob and Levitt (2003), for instance, detect teacher cheating in part by looking for poor subsequent performance of students who made rapid gains. In our data, cheating is unlikely both because our test is relatively low stakes and because our project staff administered the exam directly to avoid this possibility. Third, critics of high stakes testing often argue that shallow “teaching to the test” leads to low persistence. This is also an unlikely explanation in our context; our exam is relatively low stakes, is not part of the standard educational infrastructure, and covers only subject matter that all students should know and that Pakistani parents generally demand.¹⁵

We looked at a couple of other candidate explanations, but there are no “smoking guns” that could explain low persistence. Tables 8 and 9, for instance, present the results of a preliminary exercise that assesses whether low persistence arises from household and school responses to unexpected achievement shocks—an explanation that has different implications for cost-benefit analysis than simple forgetting.¹⁶ Parents’ perceptions of their child’s performance reacts strongly to unexpected gains in achievement, but there is only weak evidence of substi-

¹⁵On average, the children tested at the end of Grade 3 could complete two-digit addition, but not subtraction or multiplication (in mathematics); recognize simple words (but not sentences) in the vernacular (Urdu); and recognize alphabets and match simple three-letter words to pictures in English.

¹⁶We examine whether inputs adjust to unexpected achievement shocks for roughly 650 children for whom we have detailed information from a survey collected at households. As a measure of the unexpected shock, we first compute the residual from a regression of fourth grade scores on third grade scores and a host of known controls. We then test whether this residual predicts changes between fourth and fifth grade in parents’ perceptions of the child’s performance, expenditure on school, and time spent helping children on homework, being tutored, doing homework, and playing. We instrument for the subject specific residuals using the alternate subject residuals to lessen measurement error attenuation.

tution effects. School expenditures do drop slightly as do the hours spent helping the child on his or her homework. Minutes spent playing increases, but tuition also increases. While some of these responses are in the direction of substitution, they are generally not statistically significant. Given the detailed household data we obtained, it suggests that household substitution is unlikely to be a main driver behind low persistence. This may not be particularly surprising given that very low achievement suggests that children may be below parents’ desired learning levels.

Table 9 explores the possibility that fade out captures teachers targeting poorly performing students. If teachers target poorly performing children in each classroom, persistence should be lower within schools than between schools. To test this hypothesis, we estimate a basic lagged value-added model with no controls and instrument for lagged achievement using lagged differences in alternate subjects. We estimate this model using average school scores (between school specification) and child scores measured in deviations for the school average (within school specification). If anything, the persistence coefficient is lower for the between school regressions, suggesting that within school targeting is not the primary reason for low persistence.

Finally, Table 10 looks at whether heterogeneity in persistence can provide some hints about its origin (Semb and Ellis, 1994).¹⁷ To obtain the most power possible, we estimate the value-added model for specific sub-populations using the “predetermined inputs, uncorrelated effects, and conditionally stationary” based estimator (M10 of Tables 4, 5, and 6). Unfortunately, large standard errors make it difficult to find statistically different decay rates between groups. Learning in private schools seems to decay faster than learning in government schools, but the difference is not statistically significant. A similar pattern holds for richer families and children with educated parents. These results hint that learning decays faster for faster learners.

These candidate explanations do not yield a compelling story thus far; it could just be that children forget what they learned. Psychology and neuroscience provide some compelling evidence for this using laboratory experiments. Psychological research on the “curve of forgetting” dates back to Ebbinghaus’s (1885) seminal study on memorization. Rubin and Wenzel (1996) review the laboratory research spawned by this contribution. Semb and Ellis (1994) review classroom studies that test how much students remember after taking a course. Both literatures document the fragility of human memory. Cooper et al. (1996) studies the learning losses that children experience between spring and fall achievement tests. These losses are generally not as rapid as the effects we find, but the experiment is different: we estimate the depreciation with no inputs, whereas summer activities provide some stimulus, particularly for privileged

¹⁷To give one example, MacKenzie and White (1982) report that fade out for geographical knowledge was much higher for in-class exercises compared to field excursions (or, passive versus active learning.) Similarly, Rothstein (2008) finds heterogeneity in the long-run effects of teachers who produce equal short-run gains; Jacob, Lefgren and Sims (2008) estimate that teacher effects are only a third as persistent as achievement in general.

children.

5 Conclusion

In the absence of randomized studies, the value-added approach to estimating education production functions has gained momentum as a valid methodology for removing unobserved individual heterogeneity in assessing the contribution of specific programs or in understanding the contribution of school-level factors for learning (e.g. Boardman and Murnane, 1979; Hanushek, 1979; Todd and Wolpin, 2003; Hanushek, 2003; Doran and Izumi, 2004; McCaffrey, 2004; Gordon, Kane and Staiger, 2006). In such models, assumptions about learning persistence and unobserved heterogeneity play central roles. Our results reject both the assumption of perfect persistence required for the restricted value-added model and of no learning heterogeneity required for the lagged value-added model. Our results for Pakistan should illustrate the danger of incorrectly modeling or estimating education production functions: the restricted value-added model is fundamentally misspecified and can even yield wrong-signed estimates of a program’s impact. Underscoring the potential of affordable, mainstream, private schools in developing countries, we find that Pakistan’s private schools contribute roughly 0.25 standard deviations more to achievement each year than government schools, an effect greater than the average yearly gain between third and fourth grade.

Our estimate of persistence is consistent with recent work on teacher effects, with analytical and empirical estimates of the expected bias under OLS, and with experimental evidence of program fade out in developing and developed countries. But the economic interpretation still remains an open area of enquiry. Our context and test largely rule out mechanical explanations of low persistence. We also find little evidence that low persistence results from substitution by parents and teachers; the behavioral adjustments we are able to measure are unlikely to represent the primary reason achievement gains fade out. Simple forgetting, consistent with a large body of memory research in psychology, appears to be a likely explanation and hence a core component of education production functions, although more research is needed to provide direct evidence for it.

Our results also highlight that short evaluations, even when experimental, may yield little information about the cost-effectiveness of a program. Using the one or two year increase from a program gives an upper-bound on the longer term achievement gains. As our estimates suggest, and Table 7 confirms, we should expect program impacts to fade quickly. Calculating the internal rate of return by citing research linking test scores to earnings of young adults is therefore a doubtful proposition. The techniques described here, with three periods of data, can theoretically obtain a lower bound on cost-effectiveness by assuming exponential fade out. At the same time, the causes of fade out are equally important: if parents no longer need to hire

tutors or buy textbooks (the substitution interpretation of imperfect persistence), a program may be *cost*-effective even if test scores fade out.

Moving forward, empirical estimates of education production functions may benefit from further unpacking persistence. Overall, the agenda pleads for a richer model of education and for empirical techniques for modelling the broader learning process, not simply to add nuance to our understanding of learning, but to get the most basic parameters right.

A Analytical Corrections for Measurement Error

Consider the lagged value-added model

$$y_{it}^* = \alpha' \mathbf{x}_{it} + \beta y_{i,t-1}^* + v_{it}, \quad (11)$$

where y_{it}^* and $y_{i,t-1}^*$ are true achievement, v_{it} is the error term, and we have put aside the possibility of omitted heterogeneity. Since achievement is a latent variable, we can only estimate it with error. Thus, we actually estimate

$$y_{it} = \alpha' \mathbf{x}_{it} + \beta y_{i,t-1} + [v_{it} + \varepsilon_{it} - \beta \varepsilon_{i,t-1}] \quad (12)$$

and OLS is inconsistent because $y_{i,t-1}$ is correlated with $\varepsilon_{i,t-1}$.

The analytic correction we apply replaces $y_{i,t-1}$ with the best linear predictor

$$\tilde{y}_{i,t-1} \equiv \mathbb{E}^*[y_{i,t-1}^* \mid y_{i,t-1}, \mathbf{x}_{it}] = \lambda' \mathbf{x}_{it} + r_{i,t-1} y_{i,t-1}, \quad (13)$$

where λ and $r_{i,t-1}$ are parameters. To see why this works, add and subtract $\beta \tilde{y}_{i,t-1}$ from (11) to get

$$y_{it} = \mathbf{x}_{it} \alpha + \beta \tilde{y}_{i,t-1} + [\beta(y_{i,t-1} - \tilde{y}_{i,t-1}) + v_{it} + \varepsilon_{it} - \beta \varepsilon_{i,t-1}] \quad (14)$$

$$= \mathbf{x}_{it} \alpha + \beta \tilde{y}_{i,t-1} + [\beta(y_{i,t-1}^* - \tilde{y}_{i,t-1}) + v_{it} + \varepsilon_{it}]. \quad (15)$$

where the second line follows from $y_{i,t-1}^* = y_{i,t-1} - \varepsilon_{i,t-1}$. Assuming exogeneity with respect to $v_{it} + \varepsilon_{it}$, OLS is consistent if

$$\mathbb{E}[\mathbf{x}_{it}'(y_{i,t-1}^* - \tilde{y}_{i,t-1})] = 0, \quad (16)$$

$$\mathbb{E}[\tilde{y}_{i,t-1}(y_{i,t-1}^* - \tilde{y}_{i,t-1})] = 0. \quad (17)$$

These conditions are automatically satisfied since the fitted value $\tilde{y}_{i,t-1}$ and independent variables \mathbf{x}_{it} are orthogonal to the residual $y_{i,t-1}^* - \tilde{y}_{i,t-1}$ by the definition of the projection (13).

The only difficulty is estimating the projection parameters λ and $r_{i,t-1}$ since the dependent variable $y_{i,t-1}^*$ is unobserved. But it turns out that we do not need to observe the true score. The orthogonality conditions that define the projection (13) are

$$\mathbb{E}[\mathbf{x}_{it}'(y_{i,t-1}^* - \lambda' \mathbf{x}_{it} - r_{i,t-1} y_{i,t-1})] = 0, \quad (18)$$

$$\mathbb{E}[y_{i,t-1}'(y_{i,t-1}^* - \lambda' \mathbf{x}_{it} - r_{i,t-1} y_{i,t-1})] = 0. \quad (19)$$

Solving first for λ , we have

$$\lambda = \mathbb{E}[\mathbf{x}_{it}' \mathbf{x}_{it}]^{-1} \mathbb{E}[\mathbf{x}_{it}'(y_{i,t-1}^* - r_{i,t-1} y_{i,t-1})]. \quad (20)$$

Plugging (20) into (19) and solving for $r_{i,t-1}$ yields

$$r_{i,t-1} = \mathbb{E}[y_{i,t-1} \mathbf{m}_x y_{i,t-1}]^{-1} \mathbb{E}[y_{i,t-1} \mathbf{m}_x y_{i,t-1}^*] \quad (21)$$

$$= \mathbb{E}[e_{i,t-1}^2]^{-1} (\mathbb{E}[e_{i,t-1}^2] - \mathbb{E}[\varepsilon_{i,t-1}^2]) \quad (22)$$

$$= \frac{\sigma_{e_{i,t-1}}^2 - \sigma_{\varepsilon_{i,t-1}}^2}{\sigma_{e_{i,t-1}}^2}, \quad (23)$$

where $\mathbf{m}_x \equiv 1 - \mathbf{x}_{it}'(\mathbf{x}_{it}'\mathbf{x}_{it})^{-1}\mathbf{x}_{it}'$ is an annihilator vector and $e_{i,t-1}$ is the residual from a regression of $y_{i,t-1}$ on \mathbf{x}_{it} . We can estimate $r_{i,t-1}$ by computing $\sigma_{e_{i,t-1}}^2$ from the regression of $y_{i,t-1}$

on \mathbf{x}_{it} and taking $\sigma_{\varepsilon_{i,t-1}}^2$ from IRT—i.e., from the inverse Fisher information matrix of $\hat{\theta}$. Intuitively, $r_{i,t-1}$ is the heteroscedastic reliability ratio of the score minus the variation explained by the independent variables. That is, the reliability ratio of $y_{i,t-1} - \mathbb{E}[y_{i,t-1}^* | \mathbf{x}_{it}]$.

We compute λ by plugging $r_{i,t-1}$ into (20) to get

$$\lambda = \mathbb{E}[\mathbf{x}_{it}'\mathbf{x}_{it}]^{-1} \mathbb{E}[\mathbf{x}_{it}'(y_{i,t-1}^* - r_{i,t-1}y_{i,t-1})] \quad (24)$$

$$= \mathbb{E}[\mathbf{x}_{it}'\mathbf{x}_{it}]^{-1} \mathbb{E}[\mathbf{x}_{it}'y_{i,t-1}](1 - r_{i,t-1}). \quad (25)$$

The best predictor is

$$\tilde{y}_{i,t-1} = \mathbb{E}[y_{i,t-1} | \mathbf{x}_{it}](1 - r_{i,t-1}) + r_{i,t-1}y_{i,t-1} \quad (26)$$

This takes the familiar form of an empirical Bayes estimate that shrinks the observed score to the predicted mean. The shrinkage performs the same function as blowing up the coefficient using the reliability ratio after estimation. Here, however, our shrunk estimate provides a more efficient correction by using the full heteroscedastic error structure (Sullivan, 2001).

Table A1 reports persistence coefficients corrected only for measurement error using the instrumental variable (using alternate subjects) and the analytical correction approach. Each cell contains the estimated coefficient on lagged achievement from a regression with no controls and the associated standard error. Where applicable, we also report the p-value for Hansen's overidentification test statistic. This is possible for the instrumental variables estimators since we have three subject tests and three years of data.

Absent any correction (OLS), the estimated persistence coefficient ranges between 0.65 and 0.70. Instrumenting using alternate subjects raises the estimated coefficient significantly to 0.85 for English, 0.89 for mathematics, and 0.97 for Urdu. However, the overidentifying restriction is rejected at the one percent level in all three subjects. This suggests that measurement errors may be correlated across subjects at the same sitting and that this correlation may differ depending on the subject. By comparison, when we instrument for lagged achievement using double lagged scores we cannot reject the overidentifying restrictions. Unfortunately, in the context of dynamic panel methods, additional lags to address measurement error require $T = 4$. The final line of Table A1 shows estimates based on our analytical correction around 0.9. Of course, all of these estimates remain biased upward by learning heterogeneity.

References

- Alderman, H., J. Kim and P.F. Orazem. 2003. "Design, Evaluation, and Sustainability of Private Schools for the Poor: The Pakistan Urban and Rural Fellowship School Experiments." *Economics of Education Review* 22(3):265–274.
- Alderman, H., P.F. Orazem and E.M. Paterno. 2001. "School Quality, School Cost, and the Public/Private School Choices of Low-Income Households in Pakistan." *The Journal of Human Resources* 36(2):304–326.
- Altonji, J.G., T.E. Elder and C.R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113(1):151–184.
- Andrabi, T., J. Das and A.I. Khwaja. 2008. "A dime a day: The possibilities and limits of private schooling in Pakistan." *Comparative Education Review* 52(3):329–355.
- Andrabi, Tahir, Jishnu Das and Asim Ijaz Khwaja. 2006. "A dime a day : the possibilities and limits of private schooling in Pakistan." *World Bank Policy Research Working Paper 4066*.
- Angrist, J., E. Bettinger, E. Bloom, E. King and M. Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *The American Economic Review* 92(5):1535–1558.
- Arellano, M. 2003. *Panel Data Econometrics*. Oxford University Press.
- Arellano, M. and O. Bover. 1995. "Another look at the instrumental variable estimation of error-components models." *Journal of Econometrics* 68(1):29–51.
- Arellano, M. and S. Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58(2):277–297.
- Arellano, Manuel and Bo Honore. 2001. Panel data models: some recent developments. In *Handbook of Econometrics*, ed. J.J. Heckman and E.E. Leamer. Vol. 5 of *Handbook of Econometrics* Elsevier chapter 53, pp. 3229–3296.
- Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122(3).
- Blundell, R. and S. Bond. 1998. "Initial conditions and Moment Conditions in Dynamic Panel Data Models." *Journal of Econometrics* 87(1):115–43.
- Boardman, A.E. and R.J. Murnane. 1979. "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." *Sociology of Education* 52(2):113–121.
- Chay, K.Y., P.J. McEwan and M. Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools." *The American Economic Review* 95(4):1237–1258.

- Cooper, H., B. Nye, K. Charlton, J. Lindsay and S. Greathouse. 1996. "The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review." *Review of Educational Research* 66(3):227–68.
- Cunha, F. and J.J. Heckman. 2007. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* .
- Cunha, F, JJ Heckman and SM Schennach. 2006. "Estimating the Elasticity of Substitution Between Early and Late Investments in the Technology of Cognitive and Noncognitive Skill Formation." *Unpublished, University of Chicago, Department of Economics* .
- Currie, J. and D. Thomas. 1995. "Does Head Start Make a Difference?" *The American Economic Review* 85(3):341–364.
- Deming, David. 2008. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." Harvard University. Processed.
- Doran, H. and L.T. Izumi. 2004. "Putting Education to the Test: A Value-Added Model for California." *San Francisco: Pacific Research Institute* .
- Ebbinghaus, H. 1885. *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University.
- Glewwe, P., N. Ilias and M. Kremer. 2003. "Teacher Incentives." *NBER Working Paper* .
- Gordon, Robert, Thomas J. Kane and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." *Hamilton Project Discussion Paper* .
- Hanushek, E.A. 1979. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions." *The Journal of Human Resources* 14(3):351–388.
- Hanushek, E.A. 2003. "The Failure of Input-Based Schooling Policies." *Economic Journal* 113(485):64–98.
- Harris, D. and T.R. Sass. 2006. "Value-Added Models and the Measurement of Teacher Quality." *Unpublished manuscript* .
- Jacob, B. A. and L. Lefgren. 2005. "What Do Parents Value in Education: An Empirical Investigation of Parents' Revealed Preferences for Teachers." *NBER Working Paper 11494* .
- Jacob, Brian, Lars John Lefgren and David Sims. 2008. "The Persistence of Teacher-Induced Learning Gains." *NBER Working Paper* .
- Jacob, Brian and S.D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *The Quarterly Journal of Economics* 118(3):843–877.
- Jimenez, E., M.E. Lockheed and V. Paqueo. 1991. "The relative efficiency of private and public schools in developing countries." *The World Bank Research Observer* 6(2):205–218.
- Kane, T.J. and D.O. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *The Journal of Economic Perspectives* 16(4):91–114.

- Kane, T.J. and D.O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." *Unpublished. Cambridge, MA: Harvard University* .
- Kremer, M., E. Miguel and R. Thornton. 2003. "Incentives to Learn." *NBER Working Paper* .
- Krueger, A.B. 2003. "Economic Considerations and Class Size." *Economic Journal* .
- Krueger, A.B. and D.M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project Star." *The Economic Journal* 111(468):1–28.
- Ladd, H.F. and R.P. Walsh. 2002. "Implementing value-added measures of school effectiveness: getting the incentives right." *Economics of Education Review* 21(1):1–17.
- Lord, F.M. 1967. "A paradox in the interpretation of group comparisons." *Psychological Bulletin* 68(5):304–305.
- MacKenzie, A.A. and R.T. White. 1982. "Fieldwork in Geography and Long-Term Memory Structures." *American Educational Research Journal* 19(4):623–632.
- McCaffrey, D.F. 2004. *Evaluating Value-added Models for Teacher Accountability*. Rand Corporation.
- Muralidharan, Karthik and Michael Kremer. forthcoming. Public and Private Schools in Rural India. In *School Choice International*, ed. Paul Peterson and Rajashri Chakrabarti.
- Murnane, R.J., J.B. Willett and F. Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *The Review of Economics and Statistics* 77(2):251–266.
- Neal, D. and W. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differentials." *Journal of Political Economy* 104(5):869–895.
- Rothstein, Jesse. 2008. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Working Paper* .
- Rubin, DC and AE Wenzel. 1996. "One Hundred Years Of Forgetting: A Quantitative Description Of Retention." *Psychological Review* 103(4):734–760.
- Santibanez, Lucrecia. 2006. "Why we should care if teachers get A's: Teacher test scores and student achievement in Mexico." *Economics Of Education Review* 25(5):510–520.
- Sass, T.R. 2006. "Charter Schools and Student Achievement in Florida." *Education Finance and Policy* 1(1):91–122.
- Schwartz, A.E. and J. Zabel. 2005. The Good, the Bad, and the Ugly: Measuring School Efficiency Using School Production Functions. In *Measuring School Performance and Efficiency: Implications for Practice and Research*, ed. L. Stiefel, A. E. Schwartz, R. Rubenstein and J. Zabel. NY: Eye on Education, Inc. pp. 37–66.
- Schweinhardt, L. J., J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield and M. Nores. 2005. *Lifetime effects: The High/Scope Perry Preschool study through age 40*. Ypsilanti, MI: High/Scope Press.

- Semb, G.B. and J.A. Ellis. 1994. "Knowledge taught in school: What is remembered." *Review of Educational Research* 64(2):253–286.
- Sullivan, D.G. 2001. "A Note on the Estimation of Linear Regression Models with Heteroskedastic Measurement Errors." *Federal Reserve Bank of Chicago* .
- Todd, P.E. and K.I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113(485):3–33.
- Tooley, J. and P. Dixon. 2003. *Private schools for the poor: A case study from India*. Reading, Royaume-Uni: Centre for British Teachers.

TABLE 1. DYNAMIC PANEL SPECIFICATION

Estimator Assumptions	“Difference” Instruments	“Levels” Instruments	Notes
<i>Panel A: Static Estimates</i>			
M1. No depreciation $\beta=1$ (OLS)	n/a	n/a	Assumes perfect persistence and no or uncorrelated heterogeneity.
M2. No effects, no measurement error (OLS)	n/a	n/a	Assumes no measurement error and no effects.
M3. No effects (2SLS/IV correction)	n/a	Alternate subjects	Assumes no effects and uncorrelated measurement errors across subjects
M4. No effects (HEIV/ Analytical correction)	n/a	n/a	Assumes no effects and analytical correction is valid.
<i>Panel B: Difference GMM</i>			
M5. Strictly exogenous inputs	Inputs: 1...T Score: 1...t-2	n/a	Assumes no feedback effects.
M6. Predetermined inputs	Inputs: 1...t-1 Score: 1...t-2	n/a	None (beyond those that apply to all estimators)
<i>Panel C: Levels and Difference SGMM</i>			
M7. Predetermined inputs, constantly correlated effects	Inputs: 1...t-1 Score: 1...t-2	Δ Inputs: 1..t	Assumes effects have constant correlation with inputs.
M8. Predetermined inputs, uncorrelated effects	Inputs: 1...t-1 Score: 1...t-2	Inputs: 1..t	Assumes effects are uncorrelated with inputs (random effects).
<i>Panel D: Levels GMM (Proxy Style)</i>			
M9. Predetermined inputs, constantly correlated effects, conditional stationarity	Inputs: 1...t-1 Score: 1...t-2 (<i>not used</i>)	Δ Inputs: 1..t Δ Score: t-1	Assumes effects have constant correlation with inputs and scores are conditionally mean stationary.
M10. Predetermined inputs, uncorrelated effects, conditional stationarity	Inputs: 1...t-1 Score: 1...t-2 (<i>not used</i>)	Inputs: 1..t Δ Score: t-1	Assumes effects are uncorrelated with inputs (random effects) and scores are conditionally mean stationary.

Notes: The notes columns do not include knife-edge cases such as perfectly offsetting biases. None of the dynamic panel estimators allow for serial correlation, as written. Redundant instruments in levels and differences are dropped. Panel D lists the valid difference instruments, but our application does not use them in order to demonstrate a simple, single equation estimator.

TABLE 2. BASELINE CHARACTERISTICS OF CHILDREN IN PUBLIC AND PRIVATE SCHOOLS

Variable	Private School	Public School	Difference
Panel A: Full Sample			
Age	9.58 [1.49]	9.63 [1.35]	-0.04 (0.08)
Female	0.45	0.47	-0.02 (0.03)
English score (third grade)	0.74 [0.61]	-0.23 [0.94]	0.97*** (0.05)
Urdu score (third grade)	0.52 [0.78]	-0.12 [0.98]	0.63*** (0.05)
Math score (third grade)	0.39 [0.81]	-0.07 [1.00]	0.46*** (0.05)
N	2337	5783	
Panel B: Surveyed Child Sample			
Age	9.63 [1.49]	9.72 [1.34]	-0.09 (0.08)
Female	0.47	0.48	-0.02 (0.03)
Years of schooling	3.39 [1.57]	3.75 [1.10]	-0.35*** (0.08)
Weight z-score (normalized to U.S.)	-0.75 [4.21]	-0.64 [1.71]	-0.10 (0.13)
Height z-score (normalized to U.S.)	-0.42 [3.32]	-0.22 [2.39]	-0.20 (0.13)
Number of elder brothers	0.98 [1.23]	1.34 [1.36]	-0.36*** (0.05)
Number of elder sisters	1.08 [1.27]	1.27 [1.30]	-0.19*** (0.05)
Father lives at home	0.88	0.91	-0.04*** (0.01)
Mother lives at home	0.98	0.98	0.00 (0.01)
Father educated past elementary	0.64	0.46	0.18*** (0.02)
Mother educated past elementary	0.36	0.18	0.18*** (0.02)
Asset index (PCA)	0.78 [1.50]	-0.30 [1.68]	1.08*** (0.07)
English score (third grade)	0.74 [0.62]	-0.24 [0.95]	0.99*** (0.05)
Urdu score (third grade)	0.53 [0.78]	-0.14 [0.98]	0.67*** (0.05)
Math score (third grade)	0.42 [0.80]	-0.09 [1.02]	0.51*** (0.05)
N	1374	2657	

* Significant at the 10%; ** significant at the 5%; *** significant at 1%.

Notes: Cells contain means, brackets contain standard deviations, and parentheses contain standard errors. Standard errors for the private-public difference are clustered at the school level. Sample includes only those children tested (A) and surveyed (B) in all three years.

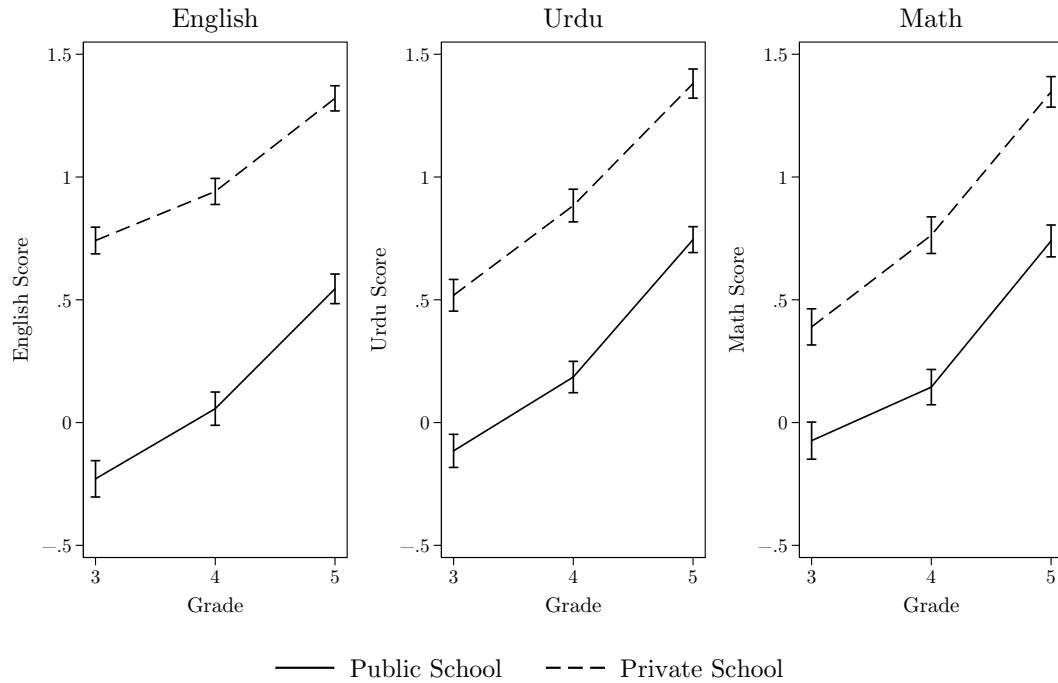
TABLE 3. THIRD GRADE ACHIEVEMENT AND CHILD, HOUSEHOLD AND SCHOOL CHARACTERISTICS

Dependent variable (third grade):	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	English	English	English	Urdu	Urdu	Urdu	Math	Math	Math
Private School	0.985 (0.047)***	0.907 (0.048)***	0.916 (0.048)***	0.670 (0.049)***	0.595 (0.050)***	0.575 (0.047)***	0.512 (0.051)***	0.446 (0.053)***	0.451 (0.052)***
Age		0.004 (0.013)	0.015 (0.012)		0.013 (0.013)	0.013 (0.012)		0.033 (0.014)**	0.048 (0.013)***
Female		0.125 (0.047)***	0.133 (0.041)***		0.209 (0.046)***	0.205 (0.040)***		-0.040 (0.051)	-0.057 (0.043)
Years of schooling		-0.029 (0.013)**	-0.019 (0.012)		-0.039 (0.014)***	-0.028 (0.014)**		-0.038 (0.015)**	-0.025 (0.014)*
Number of elder brothers		-0.030 (0.011)***	-0.035 (0.010)***		-0.020 (0.012)*	-0.025 (0.011)**		-0.020 (0.012)*	-0.023 (0.011)**
Number of elder sisters		0.008 (0.011)	0.013 (0.010)		0.001 (0.012)	-0.001 (0.012)		-0.002 (0.013)	-0.006 (0.012)
Height z-score (normalized to U.S.)		0.027 (0.007)***	0.016 (0.006)***		0.017 (0.006)***	0.012 (0.006)**		0.034 (0.008)***	0.024 (0.007)***
Weight z-score (normalized to U.S.)		-0.005 (0.008)	-0.001 (0.006)		-0.004 (0.005)	0.001 (0.005)		-0.009 (0.007)	-0.002 (0.006)
Asset index		0.041 (0.012)***	0.050 (0.009)***		0.043 (0.011)***	0.045 (0.010)***		0.030 (0.011)***	0.034 (0.010)***
Mother educated past elementary		0.048 (0.036)	0.062 (0.031)**		0.014 (0.040)	0.011 (0.035)		0.023 (0.040)	-0.006 (0.037)
Father educated past elementary		0.061 (0.033)*	0.066 (0.028)**		0.062 (0.034)*	0.049 (0.031)		0.069 (0.035)**	0.053 (0.032)*
Mother lives at home		-0.131 (0.095)	-0.025 (0.081)		-0.174 (0.102)*	-0.108 (0.092)		-0.210 (0.097)**	-0.091 (0.090)
Father lives at home		0.006 (0.049)	-0.038 (0.044)		0.019 (0.053)	0.005 (0.048)		-0.009 (0.057)	-0.026 (0.051)
Survey Date		0.003 (0.002)	0.000 (0.004)		0.001 (0.002)	0.004 (0.003)		0.003 (0.002)	0.003 (0.003)
Constant	-0.243 (0.038)***	-49.721 (38.467)	-3.690 (62.432)	-0.137 (0.035)***	-23.750 (31.915)	-59.528 (45.357)	-0.095 (0.038)**	-56.196 (35.415)	-51.248 (50.310)
Village Fixed Effects	No	No	Yes	No	No	Yes	No	No	Yes
Observations	4031	4031	4031	4031	4031	4031	4031	4031	4031
R-squared	0.23	0.25	0.37	0.11	0.13	0.25	0.06	0.08	0.21

* significant at 10%; ** significant at 5%; *** significant at 1%

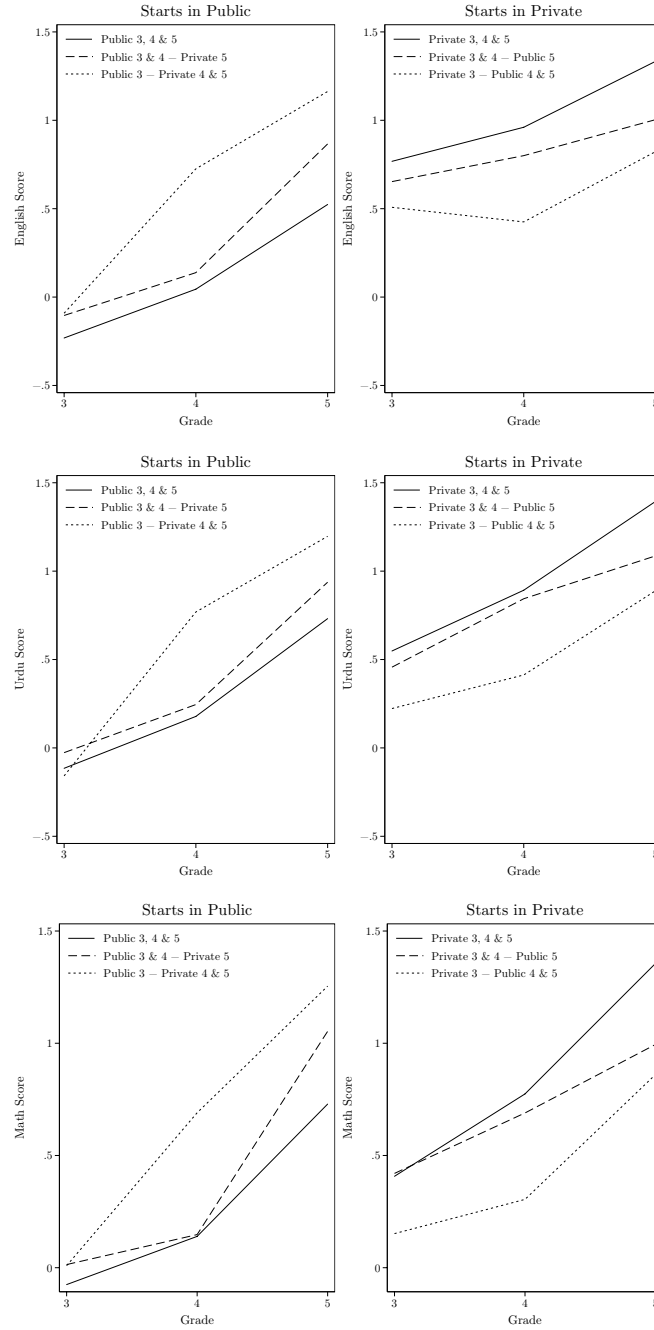
Notes: Standard errors clustered at the school level. Sample includes only those children tested and surveyed in all three years.

FIGURE 1. EVOLUTION OF TEST SCORES IN PUBLIC AND PRIVATE SCHOOLS



Notes: Vertical bars represent 95% confidence intervals around the group means, allowing for arbitrary clustering within schools. Tests scores are IRT based scale scores normalized to have mean zero and standard deviation one for the full sample of children in third grade. Children who were tested in third grade were subsequently followed and counted as being in fourth or fifth grade regardless of whether they were actually promoted. The graph's sample is limited to children who were tested in all three periods (Table 2, Panel A: Full Sample).

FIGURE 2. ACHIEVEMENT OVER TIME FOR CHILDREN WHO SWITCHED SCHOOL TYPES



	Public 3, 4, & 5	Public 3, 4 – Private 5	Public 3 – Private 4 & 5	Private 3, 4 & 5	Private 3 & 4 – Public 5	Private 3 – Public 4, 5
N	5688	40	48	2007	160	167

Notes: Lines connect group means for children who were enrolled in all three periods and have a particular private/public enrollment pattern. Children were tested in the second half of the school year; most of the gains from a child in a third grade government school and fourth grade private school should be attributed to the private school.

TABLE 4. VALUE-ADDED MODEL ESTIMATES OF PERSISTENCE AND PRIVATE SCHOOL COEFFICIENT (ENGLISH)

Estimator's Key Assumption	Persistence Coefficient	Private School Coefficient	Hansen's J	
			χ^2	df
			(p-value)	
<i>Panel A: Static Estimates</i>				
M1. No depreciation $\beta=1$ (OLS)	1.00	-0.08 (0.02)		
M2. No effects, no measurement error (OLS)	0.52 (0.02)	0.31 (0.02)		
M3. No effects (2SLS/IV Correction)	0.70 (0.02)	0.16 (0.02)	4.69 (0.03)	1
M4. No effects (HEIV/Analytical Correction)	0.74 (0.02)	0.21 (0.02)		
<i>Panel B: Difference GMM</i>				
M5. Strictly exogenous inputs	0.19 (0.10)	0.25 (0.07)	25.44 (0.02)	13
M6. Predetermined inputs	0.19 (0.10)	1.15 (0.39)	16.82 (0.02)	7
<i>Panel C: Levels and Difference SGMM</i>				
M7. Predetermined inputs, constantly correlated effects	0.36 (0.07)	0.21 (0.06)	45.50 (0.00)	23
M8. Predetermined inputs, uncorrelated effects	0.53 (0.05)	0.32 (0.04)	79.08 (0.00)	29
<i>Panel D: Levels Only GMM</i>				
M9. Predetermined inputs, constantly correlated effects, conditional stationarity	0.40 (0.05)	0.29 (0.07)	24.74 (0.02)	12
M10. Predetermined inputs, uncorrelated effects, conditional stationarity	0.39 (0.05)	0.24 (0.04)	23.43 (0.02)	11

Notes: Dots represent the estimated coefficients, thicker dark lines are 90 percent confidence intervals, and thin gray lines are 95 percent confidence intervals. All intervals and standard errors are clustered by school. See text and Table 1 for details on instruments and assumptions.

TABLE 5. VALUE-ADDED MODEL ESTIMATES OF PERSISTENCE AND PRIVATE SCHOOL COEFFICIENT (URDU)

Estimator's Key Assumption	Persistence Coefficient	Private School Coefficient	Hansen's J	
			χ^2	df
			(p-value)	
<i>Panel A: Static Estimates</i>				
M1. No depreciation $\beta=1$ (OLS)	1.00	0.01 (0.02)		
M2. No effects, no measurement error (OLS)	0.58 (0.01)	0.26 (0.02)		
M3. No effects (2SLS/IV Correction)	0.73 (0.02)	0.17 (0.02)	3.67 (0.06)	1
M4. No effects (HEIV/Analytical Correction)	0.79 (0.02)	0.20 (0.02)		
<i>Panel B: Difference GMM</i>				
M5. Strictly exogenous inputs	0.21 (0.09)	0.29 (0.07)	49.50 (0.00)	13
M6. Predetermined inputs	0.35 (0.11)	0.90 (0.48)	18.90 (0.01)	7
<i>Panel C: Levels and Difference SGMM</i>				
M7. Predetermined inputs, constantly correlated effects	0.26 (0.08)	0.22 (0.06)	66.58 (0.00)	23
M8. Predetermined inputs, uncorrelated effects	0.51 (0.06)	0.30 (0.04)	81.89 (0.00)	29
<i>Panel D: Levels Only GMM</i>				
M9. Predetermined inputs, constantly correlated effects, conditional stationarity	0.55 (0.05)	0.31 (0.07)	13.49 (0.33)	12
M10. Predetermined inputs, uncorrelated effects, conditional stationarity	0.56 (0.05)	0.27 (0.03)	13.30 (0.27)	11

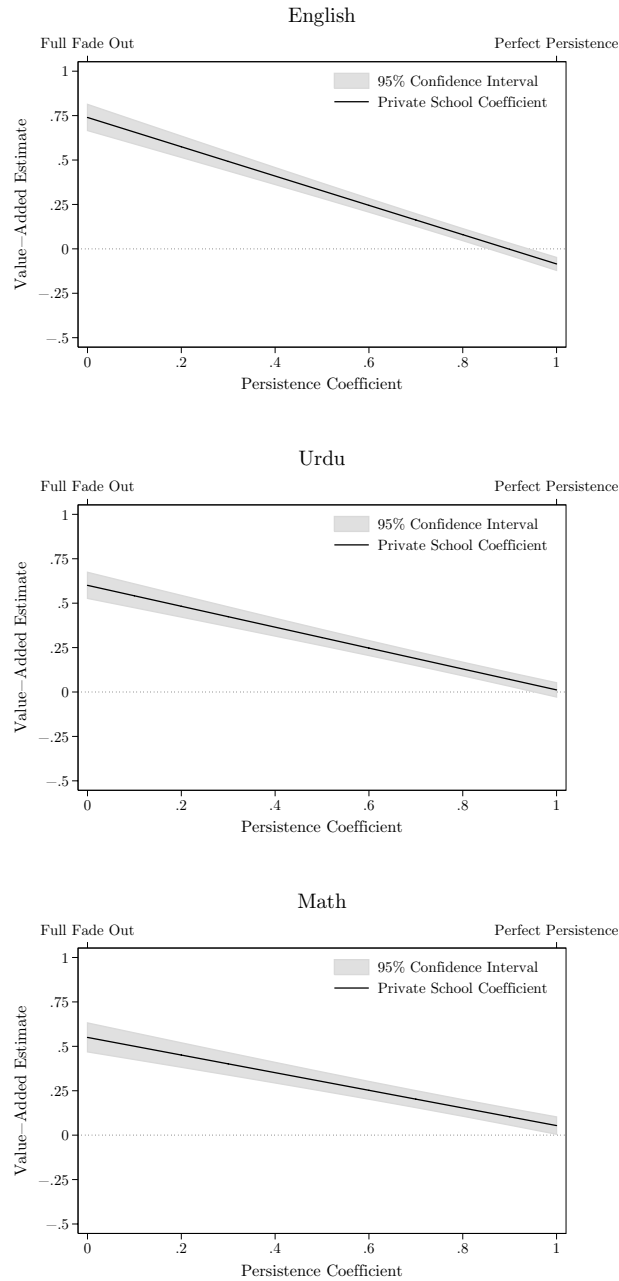
Notes: Dots represent the estimated coefficients, thicker dark lines are 90 percent confidence intervals, and thin gray lines are 95 percent confidence intervals. All intervals and standard errors are clustered by school. See text and Table 1 for details on instruments and assumptions.

TABLE 6. VALUE-ADDED MODEL ESTIMATES OF PERSISTENCE AND PRIVATE SCHOOL COEFFICIENT (MATH)

Estimator's Key Assumption	Persistence Coefficient	Private School Coefficient	Hansen's J	
			χ^2	df
			(p-value)	
<i>Panel A: Static Estimates</i>				
M1. No depreciation $\beta=1$ (OLS)	1.00	0.05 (0.02)		
M2. No effects, no measurement error (OLS)	0.57 (0.02)	0.27 (0.03)		
M3. No effects (2SLS/IV Correction)	0.76 (0.02)	0.17 (0.03)		1
M4. No effects (HEIV/Analytical Correction)	0.75 (0.02)	0.23 (0.03)		
<i>Panel B: Difference GMM</i>				
M5. Strictly exogenous inputs	-0.00 (0.09)	0.26 (0.08)		13
M6. Predetermined inputs	0.12 (0.12)	0.46 (0.50)		7
<i>Panel C: Levels and Difference SGMM</i>				
M7. Predetermined inputs, constantly correlated effects	0.12 (0.10)	0.19 (0.08)		23
M8. Predetermined inputs, uncorrelated effects	0.51 (0.08)	0.30 (0.05)		29
<i>Panel D: Levels Only GMM</i>				
M9. Predetermined inputs, constantly correlated effects, conditional stationarity	0.51 (0.06)	0.30 (0.07)		12
M10. Predetermined inputs, uncorrelated effects, conditional stationarity	0.53 (0.06)	0.27 (0.04)		11

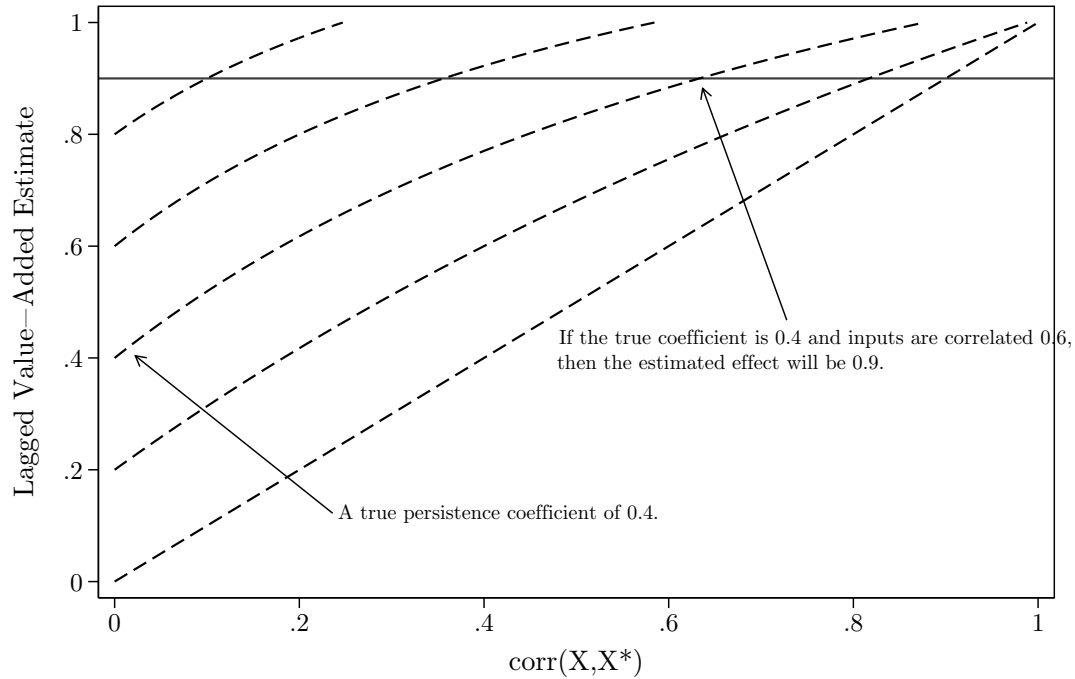
Notes: Dots represent the estimated coefficients, thicker dark lines are 90 percent confidence intervals, and thin gray lines are 95 percent confidence intervals. All intervals and standard errors are clustered by school. See text and Table 1 for details on instruments and assumptions.

FIGURE 3. PRIVATE SCHOOL VALUE-ADDED ASSUMING VARIOUS PERSISTENCE RATES



Notes: These graphs show the estimated value-added effect of private schools depending on the assumed persistence coefficient of lagged achievement. The restricted value-added model, for example, assumes the persistence coefficient equals one—no fade-out. The estimated value-added pooled for third to fourth and fourth to fifth grades is estimated by OLS controlling for age, gender, years of schooling, weight z-score, height z-score, number of elder brothers, number of elder sisters, whether father lives at home, whether mother lives at home, whether father educated past elementary, whether mother educated past elementary, an asset index, survey date, and round and village fixed effects. The confidence intervals are based on standard errors clustered at the school level.

FIGURE 4. TRUE AND ESTIMATED PERSISTENCE IN A LAGGED VALUE-ADDED MODEL WITH
SERIALLY CORRELATED OMITTED INPUTS



Notes: In the lagged value-added model, the persistence coefficient is biased upward by the correlation between omitted contemporaneous inputs and past inputs that are captured in the lagged test score. Assuming constant correlation between any two years of inputs X and X^* , the bias can be calculated analytically (see text). The graph above gives the implied bias for fourth grade. The dashed lines represent the true persistence coefficient, indicated by the associated β when $\text{corr}(X, X^*)=0$. The y-axis gives the biased estimate that results from estimating a lagged value-added model. This estimate depends on the true persistence rate (dashed lines) and on the assumed correlation of inputs over time (x-axis). For example, a (biased) estimated persistence coefficient of 0.9 may result from a true persistence coefficient of 0.4 and correlation between inputs around 0.6. These calculations that assume achievement is measured perfectly and that all inputs are omitted (i.e. unobserved) in the regression.

TABLE 7. EXPERIMENTAL ESTIMATES OF PROGRAM FADE OUT

Program	Subject	Immediate	Extended	Implied	Source
		Treatment Effect	Treatment Effect	Persistence Coefficient	
Balsakhi Program	Math	0.348	0.030	0.086	Banerjee et al (2007)
	Verbal	0.227	0.014	0.062	
CAL Program	Math	0.366	0.097	0.265	Banerjee et al (2007)
	Verbal	0.014	-0.078	~0.0	
Learning Incentives	Multi-subject	0.23	0.16	0.70	Kremer et al (2003)
Teacher Incentives	Multi-subject	0.139	-0.008	~0.0	Glewwe et al (2003)
STAR Class Size Experiment	Stanford-9 and CTBS	~5 percentile points	~2 percentile points	~ .25 to .5	Krueger and Whitmore (2001)
Summer School and	Math	0.136	0.095	0.70	Jacob and Lefgren (2004)
Grade Retention	Reading	0.104	0.062	0.60	

Notes: Extended treatment effect is achievement approximately one year after the treatment ended. Unless otherwise noted, effects are expressed in standard deviations. Results for Kremer et al. (2003) are averaged across boys and girls. Estimated effects for Jacob and Lefgren (2004) are taken for the third grade sample.

TABLE 8. HOUSEHOLD RESPONSES TO PERFORMANCE SHOCKS

Household Changes (Grade 4 to 5)	Test Score Residual (Grade 4)			N
	English	Urdu	Math	
Perception of child performance	0.25*** (0.09)	0.17** (0.07)	0.20*** (0.08)	652
Log expenditure on school	-0.05 (0.05)	-0.07* (0.04)	-0.03 (0.04)	643
Hours helping child	-0.66 (0.46)	-0.34 (0.36)	-0.44 (0.38)	645
Log minutes spent on homework	-0.14 (0.25)	0.10 (0.20)	0.04 (0.21)	617
Log minutes for tuition (tutoring)	0.21 (0.20)	0.10 (0.16)	0.02 (0.17)	620
Log minutes spent playing	0.54* (0.30)	0.27 (0.23)	0.29 (0.25)	619

* significant at 10%; ** significant at 5%; *** significant at 1%

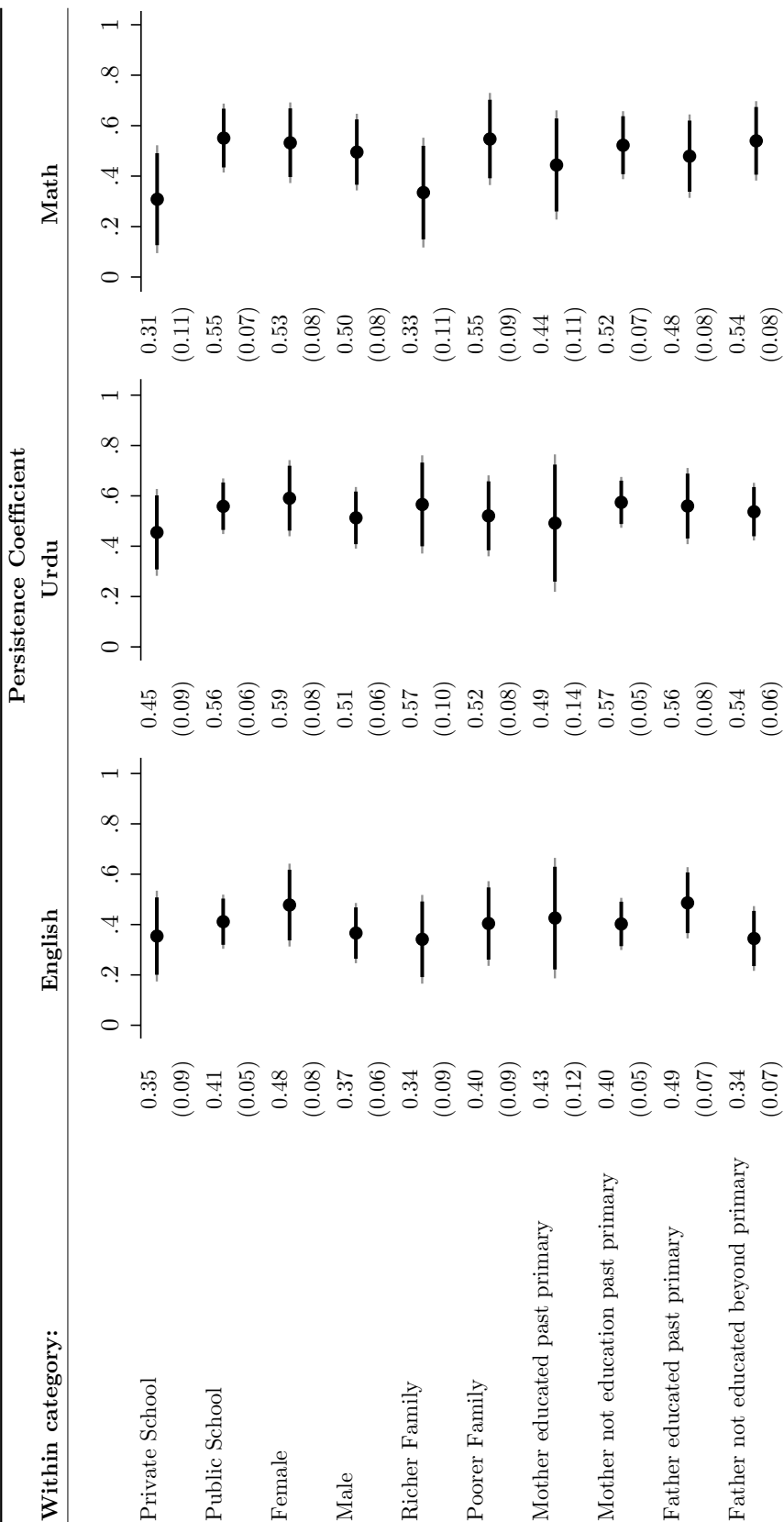
Notes: The grade 4 test score residual is computed from a lagged value-added model OLS regression that controls for third grade scores and a comprehensive set of household controls (age, gender, health status, household size, elder brothers, elder sisters, father education, mother education, adult education index, minutes spent helping child, asset index, log monthly expenditure, and wealth relative to village). The coefficients and standard errors reported are for separate 2SLS regressions of the grade 4 to 5 household behavior change on the subject residual, instrumented using the alternate subject residuals. Roughly half of the households received the score results as part of a randomized evaluation of school and child report cards. Logged variables are computed as $\ln(1+x)$.

TABLE 9. PERSISTENCE COEFFICIENT USING WITHIN AND BETWEEN SCHOOL
VARIATION ONLY

Subject	Variation	Persistence Coefficient	
			0 .2 .4 .6 .8 1
English	Within	0.57 (0.02)	
	Between	0.45 (0.09)	
Urdu	Within	0.64 (0.02)	
	Between	0.31 (0.10)	
Math	Within	0.64 (0.03)	
	Between	0.13 (0.14)	

Notes: Persistence coefficient are calculated using a 2SLS regression of test scores on lagged test scores, instrumented using lagged differences in alternate subjects (i.e. basic moments from conditional stationarity). Within regressions use school demeaned child scores whereas between regressions use mean school scores. The sample is from Table 2, Panel A with no covariates. Within N = 8620, between N = 761.

TABLE 10. PERSISTENCE COEFFICIENT HETEROGENEITY ACROSS SCHOOL, CHILD AND HOUSEHOLD CHARACTERISTICS



Notes: This table reports estimates for specific sub-populations; each coefficient is from a separate regressions. The coefficients are estimated using 2SLS (levels only) under the assumption of predetermined inputs, constantly correlated effects and conditional stationarity. Standard errors are clustered at the school level.

TABLE A1. CORRECTING FOR MEASUREMENT ERROR BIAS

Strategy	English	Urdu	Math
No Correction (OLS)	0.65 (0.015)	0.66 (0.013)	0.69 (0.013)
Alternate Subject Scores (2SLS)	0.85 (0.018) [0.000]	0.89 (0.015) [0.000]	0.97 (0.019) [0.000]
Lagged Scores (2SLS)	0.88 (0.019) [0.140]	0.86 (0.019) [0.637]	0.93 (0.020) [0.262]
Alternate Subjects and Lagged Scores (2SLS)	0.81 (0.016) [0.000]	0.80 (0.014) [0.000]	0.85 (0.015) [0.000]
Analytical Correction (HEIV OLS)	0.90 (0.020)	0.87 (0.016)	0.88 (0.017)

Notes: Cells contain coefficients from a regression of round 3 test scores on round 2 test scores—i.e. the lagged value-added model with no covariates. Parentheses contain standard errors clustered at the school level. Brackets contain the p-value for Hansen’s J statistic testing the overidentifying restrictions. The 2SLS estimates use alternate subjects or lagged scores as instruments, or both. These estimators have 1, 2 and 3 overidentifying restrictions, respectively. The analytical correction uses the score standard errors from IRT to blow-up the estimate appropriately (see Appendix A). All regressions use the same set of children.